BEST: International Journal of Management, Information Technology and Engineering (BEST: IJMITE)
ISSN 2348-0513
Vol. 3, Issue 7, Jul 2015, 1-12
© BEST Journals

A GENETIC ALGORITHM TO DETERMINE THE STRATUM BOUNDARIES USING PRPORTIONAL ALLOCATION*

MOWAFAQ MOHAMMED AL-KASSAB[1] & SALAM SAMIR DOLMAY[2]

[1]Professor, Department of Statistics and Informatics, University of Mosul, Iraq

[2]M.Sc. University of Mosul, Iraq

ABSTRACT

The stratified sampling is a method of sampling from a population. The focus will be on determine the best stratum boundaries using proportional allocation, which makes the variance less what can be, so we get more statistical precision than with simple random sampling. Assuming that the number of strata and the total sample size are predetermined. A genetic algorithm is used to obtain the stratum boundaries and the allocated sample size depending on the objective function of minimum variance of the mean of the stratified sampling. The performance of GA algorithm is compared with some methods.

KEYWORDS: Stratified Sampling, Stratum Boundaries, Proportional Allocation, Genetic Algorithm

INTRODUCTION

Stratification is the process of grouping members of the population into relatively homogenous subgroups before sampling. The strata should be mutually exclusive: every element in the population must be assigned to only one stratum. The strata should also be collectively exhaustive: no population element can be excluded (Cochran 1977), (Rao 2000), and (Orhunbilge 2000). Then random sampling is chosen within each stratum. This often improves the representativeness of the sample by reducing sampling error, and this happen when the variability within each stratum is small and the stratum means are different from one another (Cyert and Davidson 1962).

There are several methods to determine the stratum boundaries such as cumulative square root of the frequency method (cum $f^{1/2}$) (Dalenius and Hodges 1959), (cum $f^{1/3}$) (Thomsen 1976), (cum $f^{2/3}$) (AL-Kassab 1993), (cum $f^{5/6}$) (Al-Daghistani 1995), natural classes method (NCM) (Nicolini 2001), geometric approach(Gunning and Horgan 2004), random search method (Kozak 2004) and (Ekman’s rule 1959).

In this paper, we suggest a genetic algorithm (GA) approach to determine the stratum boundaries using the proportional allocation in order to minimize the variance of the estimator. The total sample size and the number of strata are predetermined and held fixed in in the application of our GA approach.

The Determination of Stratum Boundaries Using Pro Portionalal Location

Proportional allocation is a procedure for dividing a sample among the strata in a stratified sample survey. The sample survey collects data from a population in order to estimate population characteristics. A stratified sample selects separate samples from subgroups of the population, which are called “strata” and can often increase the accuracy of survey results. In order to implement stratified sampling, it is necessary to divide the population implicitly into strata before sampling. The proportional allocation is used to allocate the sample size among the strata, and it is efficient and suitable if the variances within the stratum are similar (Cyert and Davidson 1962).

The following symbols will be used throughout this paper:

Y Stratification variable

N Population size

n sample size

L number of strata

N_h Population size in stratum h. (h= 1,2, …, L)

n_h Samplesize in stratum h

W_h Weight of stratum h

$\sigma^2{}_h$ Populationvariance in the stratum h

$\bar{Y}_h$Population mean in stratum h

$\bar{y}_{st}$ Sample mean in stratified sampling

The population is divided into L homogeneous strata. The stratum size are N_1, N_2, N_L when ($N = N_1 + N_2 + + N_L$), within each stratum a simple random sample of size n_1, n_2, …, n_L is taken where ($n = n_1 + n_2 + + n_L$).

It is important to realize that the sampling is independent in the different strata.

$$\bar{y}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h, \bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$$

And $W_h = \frac{N_h}{N}$

Then

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \ldots \quad (1)$$

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W^2{}_h \frac{\sigma^2{}_h}{n_h} \quad (2)$$

(1)And (2) given by (Cochran 1977). where

$$\sigma^2{}_h = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 \quad (3)$$

In proportional allocationn_h = nW_h … (4)

Then

$$V_{prop}(\bar{y}_{st}) = \frac{1}{n} \sum_{h=1}^{L} W_h {\sigma_h}^2 \quad (5)$$

A Genetic Algorithm for Stratification

Genetic algorithms are the best ways to solve a problem for which little is known. They are a very general algorithms and so will work in any search space. All you need to know is what you need the solution to be able to do well, and a genetic algorithm will be able to create a high quality solution. Genetic algorithm uses the principles of selection and evolution to produce several solutions to a given problem. Genetic algorithm was formally introduced in the United states in the 1970s by John Holland at University of Michigan. In 1992 John Koza has used genetic algorithm to evolve programs to perform certain tasks. He called his method “genetic programming” GP.

To use a genetic algorithm, we must represent the solution to our problem as a chromosome. The genetic algorithm then creates a population of solutions and applies genetic operators such as selection, crossover, and mutation to evolve the solutions in order to find the best one(s).

Algorithm is start with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. This is motivated by hope, that the new population will be better than the old one. Solutions which are selected to form new solutions (offspring) are selected according to their fitness, the more suitable they are the more chances they have to reproduce. This process is repeated until predetermined number of iterations is reached. The best individual in the last generation becomes the solution of the problem. (Keskintürk 2007)

The Steps of Using the Genetic Algorithm

Start: Generate random initial generation.

Fitness Function: Evaluate the fitness of each chromosome.

Selection: Select the better individuals of the next generation.

Crossover: With a crossover probability, exchange the parents to form new offspring.

Mutation: With a mutation probability, mutate new offspring.

Loop: If stopping criterion is not reached go to fitness function.

Stop: Return to the best solution in current generation.

Encoding of a Chromosome

The chromosome should in some way contain information about solution which it represents. Stratification values must be encoded into chromosomes in order to solve the stratification problem of the stratum boundaries with GA. There are several types of encoding such as binary, real-valued, and integer. The most used way of encoding is a binary string which we shall use in this paper.

In binary encoding, the total number of genes in chromosomes equals the number of problem values. In each chromosome, the number of genes from the first “0” to the first “1” refers to the size of the first stratum (N_1), from “0” that comes after the first “1” to the second “1” refers to the second stratum size (N_2) and so on. Values that correspond to the indices of genes with “1” also represent the boundaries of stratum. The number of genes represented by “1” equal to the number of strata (L). The last gene must always be “1” because it represents the upper boundary of the final stratum. (Figure 1) (Keskintürk 2007)

Fitness Function

It is a function suitable for solution in the algorithm, it depends on the quality of the issue to be resolved and possibly a minimization or a maximization function for most of the issues. The fitness function calculates for each chromosome in each generation to assess its accomplish.

In our algorithm the fitness function is minimization of thevariance of the mean of the stratified sampling denoted in Eq (5). Fitness values signifies the individual’s probability of surviving in the next generation.

Selection Process

Chromosomes are selected from the population to be parents to crossover. According to Darwin's evolution theory the best ones should survive and create new offspring. According to their fitness values, selection process determines which chromosome will survive in the next generation, and the chromosome with a better fitness value have more chance to survive. The problem is how to select the chromosomes. There are many methods to select the best chromosomes (Goldberg 1989) such as proportional selection, uniform selection and roulette wheel selection, the most widely used one is roulette wheel selection.

Roulette Wheel Selection

Parents are selected according to their fitness. The better the chromosomes are the more chances to be selected. Imagine a roulette wheel where is placed all chromosomes in the populations, everyone has its bigger place according to its fitness function. (Keskintürk 2007)

Then a marble is thrown there and selects the chromosome. Chromosome with bigger fitness will be selected more times.

Crossover Process

The crossover operator is a genetic operator that combines two chromosomes (parents) to produce a new chromosome (offspring). The idea behind crossover is that the new chromosome may be better than both of the parents if it take the best characteristics from each of the parents.

The crossover point can only be determined if the both sides of the point individual have the same number of strata (the sum of 1s that will exchange must be equal). There several techniques of crossover such as single-point crossover, two-point crossover, and others. The crossover process is applied as displayed in Figure (2) “the single-point crossover” and Figure (3) “the two-point crossover”.

Mutation Process

Mutation is a genetic operator used to change genetic algorithm chromosomes to the best for the next generation. Mutation alters one or more gene values in a chromosome from its initial state. In mutation, the solution may change entirely from the previous solution by using mutation.

In GA, there are various kinds of mutation operator such as single-point mutation, random exchange mutation, etc. (Michalewicz 1992); (Nearchon 2004). (Figure 4) shows the mutation procedure. (Figure 5) shows the whole process of GA.

Numerical Applications

In this section, we will apply the proposed method and compare it with other methods to demonstrate its efficiency to find stratum boundaries, the sample is distributed with proportional allocation method, we will use some probability distributionsand standard populations data for the comparison.

Application Using Probability Distributions

In order to show the efficiency of the proposed method, we will compare our method with (Dalenius and Hodges 1959) method. The two methods are applied using the following three probability distributions:

- f (x) = 2(1-x)**, 0 < x <**
- f (x) = e^{-x}**, 0 < x <** ∞
- f (x) = xe^{-x}**, 0 < x <** ∞

Tables 1, 2, and 3 give the stratum boundaries and $V_{prop}(\bar{y}_{st})$ for these three probability distributions for 2,3,4 and 5 strata. It can be seen from tables 1, 2 and 3 that the Genetic Algorithm gives $V_{prop}(\bar{y}_{st})$less than (Dalenius and Hodges 1959) method.

Application Using Standard Populations

In this section, many populations are used for stratification with different skewness, kurtosis, mean, standard deviation and size properties, and (Table 4) shows some of the Statistical Measures of the Standard Populations. Those populations that are available in (R stratification) and (GA4Stratification) packages are used for stratification. Each populationis divided into 3, 4, 5 and 6 strata, the total sample size is 100 and the boundaries are obtained (Kozak 2004) and (Lavallée and Hidiroglou 1988) methods with random initial boundaries.

The standard populations are:

Pop1: An accounting population of debtors in an Irish firm (Debtors)

Pop2: Number of municipal employees of 284 municipalities in Sweden in 1984 (ME84)

Pop3: Simulated Data from the Monthly Retail Trade Survey of Statistics Canada (MRTS)

Pop4: Population in thousands of 284 municipalities in Sweden in 1975 (P75)

Pop5: Real estate values (in millions of kronor) according to 1984 assessment in the 284 municipalities in Sweden (REV84).

Pop6: The resources in millions of dollars of a large commercial bank in the US (US banks)

Pop7: The population in thousands of US cities in 1940 (US cities)

Pop8: The number of students in four-year US colleges in 1952-1953 (US colleges)

The variance of the mean of the stratified sampling in Eq. (5) is used in order to compare the efficiency of the proposed method with the other methods. We see that our proposed algorithm using Mat lab programming language on a PC (CPU 3.00 GHz, 3GB RAM) variance less than the other methods, also the variance values using GA decreases with increasing number of stratum, while the other methods may not achieve this property in some populations because the selected stratum boundaries do not give the optimum results (Table 6). The Genetic Algorithm needs some properties such as populations size, iteration number, crossover rate and mutation rate to stratified these standard populations as shown in (Table 5), the stratum sizes of the populations that used in this comparison shown in (Table 7).

CONCLUSIONS

The Genetic Algorithm is more efficient than the approximate methods since it give us exact solution while the approximate methods give an approximate solution.

The Genetic Algorithm is better than the other methods when it can be modified by using iterative steps.

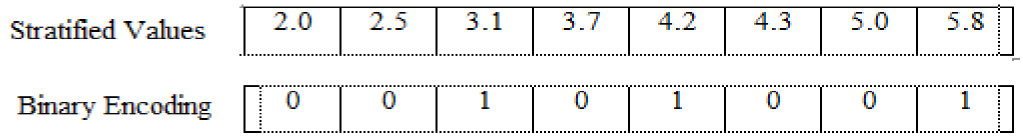
Comparing the genetic algorithm with the approximate (cum.$f^{5/6}$) method on a set of probability distributions it turns out that the genetic algorithm is more efficient than this method (tables 1,2 and 3).

When applying the genetic algorithm and the other methods on a set of standard populations it turns out that the genetic algorithm is more efficient than these methods, where found that the strata boundaries which have been found using genetic algorithm gives variance less than other methods.

We can use the Genetic Algorithm in a lot of issues that needs to divide the population into several stratato get more accurate results.

The use of evolutionary algorithms in stratified sampling, specifically in finding the optimal strata boundaries lead to an improvement in the quality of the resulting solutions because of the exploratory and exploitative capacity of these algorithms.

Figure 1: Binary Encoding of Stratification Values



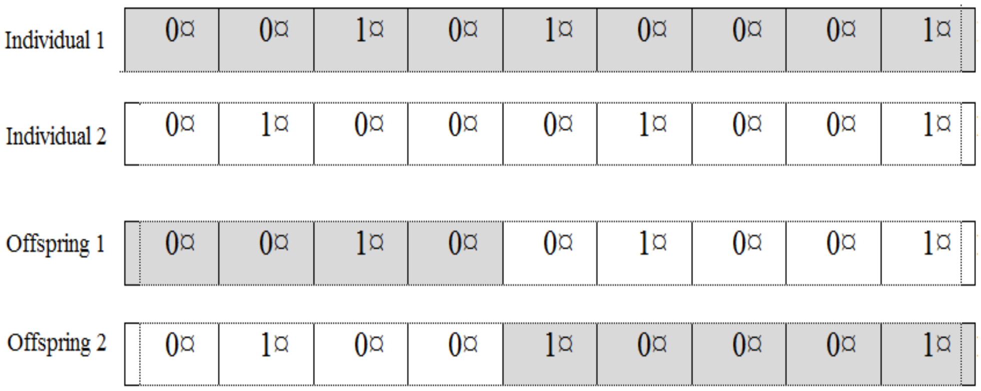
Stratified Values	2.0	2.5	3.1	3.7	4.2	4.3	5.0	5.8
Binary Encoding	0	0	1	0	1	0	0	1

Stratum 1 ⟶ 2.0 , 2.5 , 3.1 (N_1=3)

Stratum 2 ⟶ 3.7 , 4.2 (N_2=2)

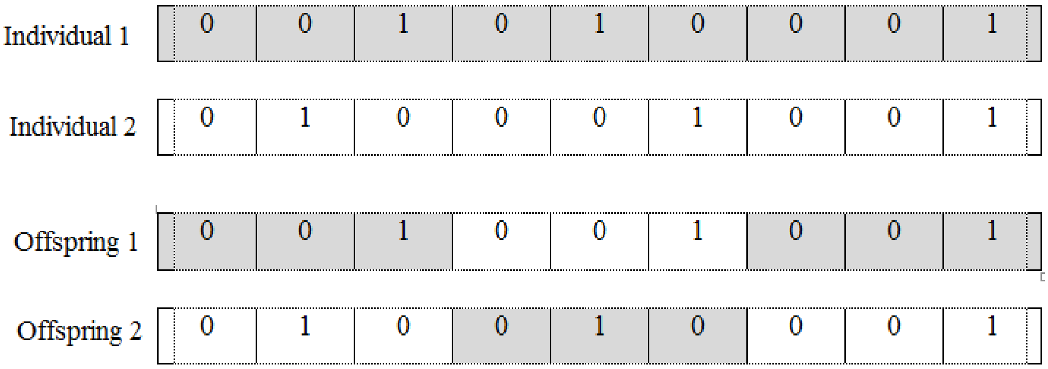
Stratum 3 ⟶ 4.3 , 5.0 , 5.8 (N_3=3)

Therefore the boundaries are 3.1, 4.2, and 5.8



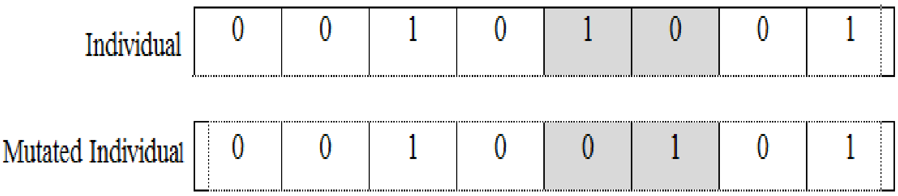
Individual 1	0¤	0¤	1¤	0¤	1¤	0¤	0¤	0¤	1¤
Individual 2	0¤	1¤	0¤	0¤	0¤	1¤	0¤	0¤	1¤
Offspring 1	0¤	0¤	1¤	0¤	0¤	1¤	0¤	0¤	1¤
Offspring 2	0¤	1¤	0¤	0¤	1¤	0¤	0¤	0¤	1¤

Figure 2: Single-Point Crossover



Individual 1	0	0	1	0	1	0	0	0	1
Individual 2	0	1	0	0	0	1	0	0	1
Offspring 1	0	0	1	0	0	1	0	0	1
Offspring 2	0	1	0	0	1	0	0	0	1

Figure 3: Two-Point Crossover



Individual	0	0	1	0	1	0	0	1
Mutated Individual	0	0	1	0	0	1	0	1

Figure 4: Mutation Procedure

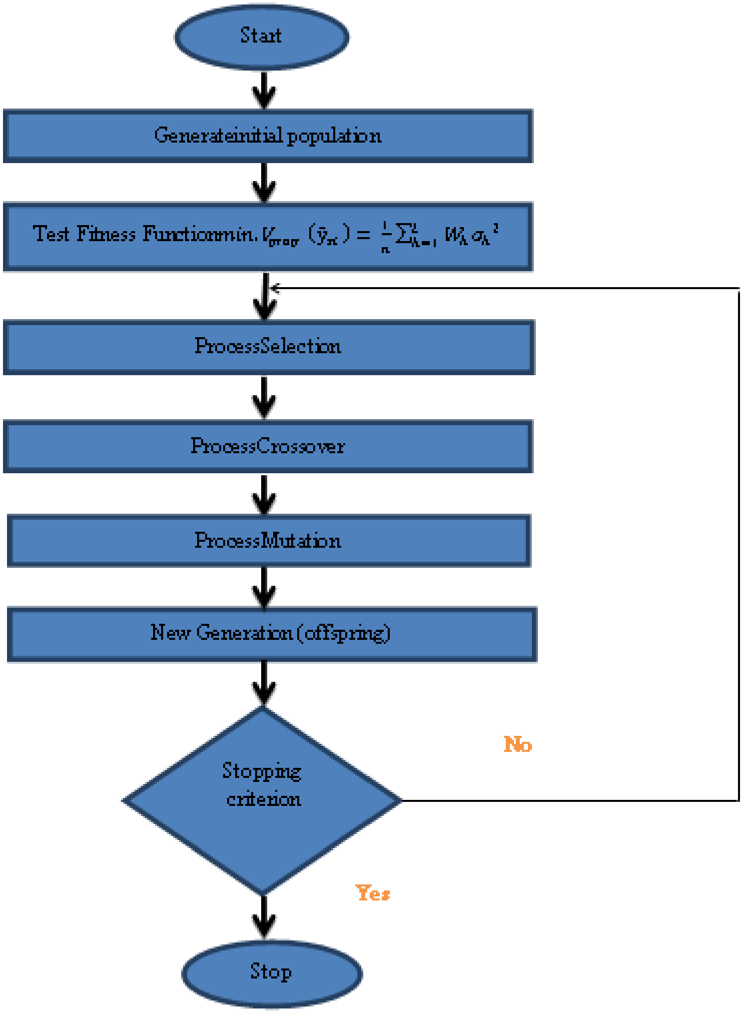

Start
Generateinitial population
Test Fitness Functionmin. $V_{prop}(\bar{y}_{st}) = \frac{1}{n}\sum_{h=1}^{L} W_h \sigma_h^2$
ProcessSelection
ProcessCrossover
ProcessMutation
New Generation (offspring)
Stopping criterion
No
Yes
Stop

Figure 5: Shows the Whole Process of GA

Table 1: Stratum Boundaries and Variance of F(X) = 2(1-X)

Number of Stratum (L)	Genetic Algorithm GA		Dalenius & Hodges (1959)	
	Strata Boundaries	Vprop $(\bar{y}_{st})$	Strata Boundaries	Vprop $(\bar{y}_{st})$
2	0.3819	0.0154	0.35	0.0157
3	0.2513	0.0071	0.23	0.073
	0.5373		0.50	
4	0.1889	0.0041	0.18	0.042
	0.3943		0.37	
	0.6259		0.62	
5	0.1531	0.0026	0.12	0.030
	0.3107		0.25	
	0.4947		0.40	
	0.6860		0.64	

Table 2: Stratum Boundaries and Variance of F(X) =e^{-x}.

Number of Stratum (L)	Genetic Algorithm GA		Dalenius & Hodges (1959)	
	Strata Boundaries	Vprop ($\bar{y}_{St}$)	Strata Boundaries	Vprop ($\bar{y}_{St}$)
2	1.5911	0.3491	1.27	0.3667
3	1.0152	0.1772	0.37	0.1958
	2.6027		2.04	
4	0.7531	0.1069	0.52	0.1227
	1.7674		1.27	
	3.3508		2.61	
5	0.6202	0.0714	0.39	0.0858
	1.3847		0.92	
	2.4173		1.68	
	4.0114		3.02	

Table 3: Stratum Boundaries and Variance of F(X) =X e^{-x}

Number of Stratum (L)	Genetic Algorithm GA		Dalenius & Hodges (1959)	
	Strata Boundaries	Vprop ($\bar{y}_{St}$)	Strata Boundaries	Vprop ($\bar{y}_{St}$)
2	2.5717	0.6863	2.36	0.6950
3	1.7884	0.3479	1.54	0.3658
	3.7613		3.26	
4	1.4078	0.2094	1.20	0.2257
	2.7254		2.27	
	4.5885		3.94	
5	1.1842	0.1395	1.01	0.1519
	2.1939		1.82	
	3.4310		2.86	
	5.2134		4.49	

Table 4: Statistical Measures of the Standard Populations

Pop	Name	N	Range	Skewness	Kurtosis	Mean	Stddev.
1	Debtors	3369	40 -28000	6.44	59.00	838.64	1873.99
2	ME84	284	173 – 47074	8.64	84.04	1779.07	4253.13
3	MRTS	2000	141 – 486366	8.61	136.20	16882.8	21574.88
4	P75	284	4 – 671	8.43	88.56	28.81	52.87
5	REV84	284	347 – 59877	7.83	81.33	3088.09	4746.16
6	USbanks	357	70 – 977	2.07	4.06	225.62	190.46
7	UScities	1038	10 – 198	2.87	9.12	32.57	30.4
8	UScolleges	677	200 - 9623	2.45	5.80	1563	1799.06

Table 5: GA Parameters Settings

GA Parameters	L=2,3,4	L=4,5
Population Size	100	150
Iteration	100	200
Crossover Rate	0.8	0.9
Mutation Rate	0.1	0.05

Table 6: The Stratum Boundaries and the Variances of the Stratified Sampling Mean (Vprop ($\bar{Y}_{st}$)) Using G A, Kozak and L&H Methods

L	Genetic Algorithm (GA)	Kozak	Lavallée and Hidiroglou (L&H)
Pop(1): Debtors			
3	6703.7	10155	7810.8
4	4135.8	7293.6	5598.5
5	2610.8	4892.6	3837.4
6	1868.5	4550	3725.2
Pop(2): ME84			
3	19525	144310	144310

Table 6: Contd.,

4	14934	144110	144530
5	12488	145310	145310
6	11394	145680	145280
Pop(3): MRTS			
3	1397000	1991000	1469500
4	906640	1229000	1181600
5	621460	1133600	1094400
6	510150	1094400	1096500
Pop(4): P75			
3	4.476515	20.287	20.747
4	3.270845	20.213	20.689
5	2.822907	20.856	20.656
6	2.652325	20.915	20.267
Pop(5): REV84			
3	35702	152900	155080
4	25989	155070	154510
5	21600	154180	154210
6	19885	154060	144820
Pop(6): USbanks			
3	39.302	72.540	71.382
4	20.955	68.339	71.276
5	14.378	62.686	70.317
6	11.413	62.233	62.233
Pop(7): UScities			
3	1.114905	1.757414	1.241580
4	0.650937	1.506219	0.885568
5	0.424536	1.463000	0.853686
6	0.305801	0.447133	0.814608
Pop(8): UScolleges			
3	3573.3	4225.9	3670.4
4	2015.5	3686.6	3216.2
5	1334.2	3506.7	3344.3
6	953.98	3147.9	3454.5

REFERENCES

1. **Al-Daghistani, T. H. N. (1995)**. An Approximately Optimal stratification using proportional allocation. Master thesis, Department of Statistics, University of Mosul, Iraq.
2. **Al-Kassab, M. M. T. (1993).** Approximately optimal stratification using proportional allocation. Journey of Tanmiat Al-Rafidain.
3. **Cochran, W.G. (1997).** Sampling Techniques. Third ed., Wiley, USA, pp. 89-101.
4. **Cyert, R. M. and Davidson, H.J. (1962).** Statistical sampling for accounting information. Prentice-Hall, Englewood Cliffs, NJ, pp. 116-127.
5. **Dalenius, Tore and Hodges, Joseph L. Jr. (1959).** Minimum variance stratification. JASA, 54(285), pp. 88-101.
6. **Ekman, G. (1959).** An Approximately useful in univariate stratification. Institute of Mathematical Statistics,30, 219-229.
7. **Goldberg, D.E. (1989).** Genetic Algorithms in search optimization and machine learning. Addison Wesley , New York , pp. 60-76.

8. **Gunning, P. and Horgan, J.M. (2004).** A new algorithm for the construction of stratum boundaries in skewed populations. Survey methodology, 30(2), pp. 159-166.

9. **Keskintürk, T. and Er, Ş. (2007).** A Genetic Algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. Computational statistics and data analysis, 52, 1, pp. 53-67.

10. **Kozak, M. (2004).** Optimal stratification using random search method in agricultural surveys. Statistics In Transition 6(5), pp. 797-806.

11. **Lavallée, P. and Hidiroglou, M. (1988).** On the stratification of skewed populations. Survey Methodology, 14,1, pp. 33-43.

12. **Michalewicz, Z. (1992).** Genetic Algorithms + Data structure = Evolution Programs. springer, Berlin.

13. **Nearchou, A.C. (2004).** The effect of various operators on the genetic search for large scheduling problems. International Journal of production Economics 88(2), pp. 191-203.

14. **Nicolini, G. (2001).** A method to define strata boundaries. Departmental working paper, 2001-01 , Department of Economics , University of Milan , Italy.

15. **Orhunbilge, N. (2000).** Sampling methods and Hypothesis tests. Second ed. Arciol Basim Yayin , Istanbul , Turkey.

16. **Rao, P.S.R.S. (2000).** Sampling methodologies with applications. Chapman & Hall/CRC press, Washington DC.

17. **R: GA4Stratification.** http://CRAN.R-project.org/package=GA4Stratification

18. **R: stratification.** http://CRAN.R-project.org/package=stratification

19. **Thomsen, I. (1976).** A comparison of Approximately Optimal stratification given proportional Allocation with other methods of stratification and Allocation. Metrika , Volume 23, pp. 15-25.

APPENDICES

Table 7: Stratum Sizes (N_h)

L	Genetic Algorithm (Ga)	Kozak	Lavallée and Hidiroglou (L&H)
Pop(1): Debtors			
3	3098 239 32	2673 561 135	2894 449 26
4	2784 445 121 19	2071 914 303 81	2179 891 271 28
5	2880 335 109 33 12	1892 954 335 139 49	1856991 350 146 26
6	25954911936715 8	1533 905 493 26512647	1608 956423 223 127 32
Pop(2): ME84			
3	254 284	145 78 61	144 79 61
4	236 42 4 2	115 64 44 61	115 62 45 62
5	190 64 25 3 2	54 69 56 41 64	54 69 54 43 64
6	177 63 23 15 4 2	54 61 33 34 37 65	42 72 32 36 38 64
Pop(3): MRTS			
3	1764 217 19	1204 688 108	1546 426 28
4	1701 269 28 2	1017 748 203 32	1017 749 206 28
5	1319 542 117 20 2	774 675 369 150 32	749 690 379 153 29
6	1106 698 161 22 4 2	513 580 458 281 136 32	513 580 455 280 140 32
Pop(4): P75			
3	248 34 2	150 77 57	132 89 63
4	219 51 12 2	111 73 43 57	64 91 66 63
5	179 68 26 9 2	64 68 52 34 66	45 66 65 45 63
6	175 59 29 14 5 2	45 66 39 34 33 67	45 34 53 52 42 58

Table 7: Contd.,

Pop(5): REV84			
3	233 49 2	138 81 65	131 84 69
4	214 55 13 2	64 81 69 70	64 81 70 69
5	158 76 36 12 2	61 69 51 34 69	61 60 47 47 69
6	138 73 35 24 12 2	57 51 37 42 28 69	50 55 40 46 39 54
Pop(6): USbanks			
3	2676822	2128461	2128560
4	22375 41 18	1111127361	1101087663
5	18877 39 3518	110101543260	70 6885 71 63
6	214573533 8 10	51639754 3260	546097 54 32 60
Pop(7): UScities			
3	859 142 37	749 193 96	795 206 37
4	775 161 70 32	434 356 154 94	393 433 173 39
5	458 364 120 65 31	226 271 298 149 94	189 270 367 171 41
6	434 356 122 7226 28	226271 285 128 89 39	154 154 271 267 14547
Pop(8): UScolleges			
3	51710753	47813069	485137 55
4	4201183633	25623411869	256242118 61
5	3681886432 25	192 166 14510569	135 20116710866
6	250237854733 25	133 179 166775369	93 151134126 10469